Effective electron density map improvement and structure validation on a Linux multi-CPU web cluster: The TB Structural Genomics Consortium Bias Removal Web Service.

Vinod Reddy [1], Stan Swanson [1], James C. Sacchettini [1], Katherine A. Kantardjieff [2], Brent Segelke [3] and Bernhard Rupp[1,3*]

[1] Biochemistry and Biophysics Department, Texas A&M University, 2128 TAMU, College Station, TX 77843-2128

[2] W.M. Keck Foundation Center for Molecular Structure, Department of Chemistry and Biochemistry, California State University Fullerton, 800 N. State College Blvd., Fullerton, CA 92834-6866

[3] Macromolecular Crystallography and Structural Genomics Group, Lawrence Livermore National Laboratory, University of California, Livermore, CA 94551

* Corresponding author: Bernhard Rupp, br@llnl.gov, Phone (925) 209–7429, Fax (801) 880-3982

RUNNING TITLE: The TB Structural Genomics Bias Removal Web Service

Synopsis: A public web service for effective map improvement, bias removal, and structure validation has been implemented on a LINUX multi-CPU cluster.

Keywords: Electron density improvement, bias removal, real space correlation, model validation

Abstract

Anticipating a continuing increase in the number of structures solved by molecular replacement in high throughput crystallography and drug discovery programs, we have implemented a user-friendly web service for automated molecular replacement, map improvement, bias removal, and real space correlation structure validation. The service is based on an efficient bias removal protocol, *Shake&wARP*, and implemented using *EPMR* and the *CCP4* suite of programs, combined with various shell scripts and FORTRAN90 routines. The service returns improved maps, converted data files, and real space correlation and B-factor plots. User data are uploaded though a web interface, and the CPU intensive iteration cycles executed on a low-cost, LINUX multi CPU cluster using the Condor job queuing package. Examples for map improvement at various resolutions are provided, and include model completion and reconstruction of absent parts, sequence correction, and ligand validation in drug target structures.

Introduction

The number of structures obtained by molecular replacement is expected to grow rapidly in the coming years both in academic labs and in pharmaceutical structure based drug discovery efforts (Blundell et al., 2001). Increasing accessibility of powerful molecular replacement programs, an increasing availability of search models due to the discovery of novel folds by public and commercial structural genomics efforts (Norvell and Zapp-Machalek, 2000) are major contributing factors. An estimate of structures solved in a commercial structural genomics effort indicates that about 70% of all structures processed were solved by MR, and in drug discovery efforts the numbers may even be higher (Kissinger et al., 2001).

Anticipating a corresponding need for map improvement and electron density based structure validation, we present an effective, easy to use web service for map improvement, phase bias removal, and rapid assessment of local model quality, which complements the geometry based structure validation programs such as *WHAT-IF*, (Vriend, 1990) and *PROCHECK* (Laskowski et al., 1993). The protocol, *Shake&wARP*, achieves effective bias removal using a modified *wARP* (Perrakis et al., 1997) procedure, and is implemented using the *CCP4* suite (CCP4, 1994) of programs, various shell scripts, and FORTRAN90 applets, executable in parallel mode on multiple CPUs. *Shake&wARP* differs in details and choice of parameters from the routines distributed with *wARP*, most notably in model perturbation, dummy atom placement /removal criteria, and map averaging. Given a modest model, *Shake&wARP* works efficiently at resolutions as low as 2.6 to 2.8Å, and it yields improved map quality in direct comparison

with other bias-reduced map reconstruction methods. The web service allows for fully automated molecular replacement (MR) solution using *EPMR* (Kissinger et al., 1999) followed by *Shake&wARP* map averaging. Residue-by-residue real space correlation coefficient and B-factor plots are automatically created in GIF format. Inspection of real space correlation coefficient plots provides a quick assessment of local structure quality, ensuring that no details have been overlooked in important areas, while less time is wasted on over-refinement in areas of little interest, and which may serve only to create artificially low global quality descriptors.

We present details of protocol and implementation of a web service, which extends our low-cost approach to high throughput protein crystallography (Rupp, 2003) to parallel execution and job queuing using Intel/AMD based hardware and Linux/Condor as portable operating system platform. A number of general examples demonstrate commonly observed phase bias and map interpretation problems and illustrate the importance of effective bias removal and electron density map improvement.

Model phase bias and map improvement

Model phase bias is a major concern in any crystallographic structure determination, in particular when the experimental phases are sub-optimal, or significantly biased, as in maps derived from molecular replacement (MR) phases. The effects of insidious model bias can be dramatic and are not easily recognized by commonly used global structure quality descriptors such as R and freeR (for a review, see (Kleywegt and Jones, 1997)). In severe cases, model bias can introduce artifacts that seriously limit the usefulness of a structure, and questionable conclusions affecting the

biological significance of results may be drawn (Rupp and Segelke, 2001, Hanson and Stevens, 2000, Hanson et al., 2002).

Source of model phase bias. Model (or phase) bias results from the fact that reflection phase angles (a_{hkl}), which are required to complete the Fourier transformation of structure amplitudes $|F|_{hkl}$ back to electron density $r_{(xyz)}$ (Equation 1) are not directly observable quantities. They must be provided by additional phasing experiments, for example, Multiple Isomorphous Replacement methods (Blundell and Johnson, 1976, Islam et al., 1998) or anomalous phasing, such as Single- (Matthews, 2001) or Multi- (Hendrickson and Ogata, 1997) wavelength Anomalous Diffraction. Phases may also originate from initial molecular replacement (MR) models, where they tend to be marginal and highly biased (Adams et al., 1999).

$$r_{(xyz)} = \frac{1}{V} \sum_{-h}^{h} \sum_{-k}^{k} \sum_{-l}^{l} |F|_{hkl} e^{-2\pi i(hx+ky+lz-a_{hkl})} \qquad (1)$$

Model Bias Reduction. The need for countermeasures against model bias has long been recognized, and a variety of bias reduction methods are presently used and implemented in modern program packages (see for example (Read, 1997) for a comprehensive discussion of model bias). A number of general strategies, or combinations thereof, are commonly employed to combat model bias: a) omission of parts of the model, b) perturbation ('shaking') of the model coordinates between refinement/rebuilding cycles, c) allowance for model errors in the refinement target functions and map coefficients, d) repeated cycling of real space and reciprocal space refinement (real space fitting of the model into electron density vs. refining model

coordinates against reciprocal space structure factors), and e) map averaging techniques. During refinement, the implementation of maximum likelihood (ML) targets, as implemented in *REFMAC* (Murshudov et al., 1997) or *CNS* (Brünger et al., 1998), together with s_A-weighted map coefficients (Read, 1986) of a general form $2m|F_o|-D|F_c|\exp(ia_c)$ accounting for partial or incorrect model, produce maps with significantly reduced model bias. When these strategies are used together with strict freeR cross-validation (Brünger, 1992), relatively 'safe' crystallography should be possible. Nevertheless, in many cases a weak part of the structure, a ligand, or cofactor may need additional individual confirmation to ensure that its density is not a result of remaining model bias. In these cases, omission of the questionable model part in the phase calculation, and perturbation of the remainder of the structure to eliminate 'memory', followed by ML refinement, will yield a map of improved quality. Electron density will either confirm a disordered residue, or perhaps obliterate the hope for the presence of a ligand. 'Classical' omit maps (Bath and Cohen, 1994, Bath, 1988), s_A-omit maps (Read, 1990, Read, 1986), simulated-annealing-omit maps (Hodel et al., 1992), and shake-omit maps (Zeng et al., 1997) are commonly used for this purpose. A ML-based reciprocal space density modification method (*Prime&Switch*), which can be applied to initial experimental maps or model phased maps and appears to perform well at low resolution and with marginal models, has been recently implemented in *RESOLVE* (Terwilliger, 2000, Terwilliger, 1999).

Map improvement and bias reduction with Shake&wARP. The basic idea behind *Shake&wARP* (*S&W)* is to combine most of the available means for bias reduction in one single protocol. The strategies implemented include omitting parts (randomly atoms

and/or specific parts of the model), perturbation ('shaking') of coordinates, use of Maximum Likelihood (ML) refinement targets in *REFMAC* (Murshudov et al., 1997) iterating in multiple cycles with real space dummy atom placement using the CCP4 program *arp_waters* (Lamzin and Wilson, 1993), and finally, probably the most effective contribution to *S&W*, averaging of six maps resulting from differently perturbed starting models (Perrakis et al., 1997).

The original *wARP* procedure per default places atoms into high density peaks (3.5s) and removes atoms below 1.5s density, although the *ARP* (now *arp_waters*) program (Lamzin and Wilson, 1993) will add additional atoms below this level, if it has not found the number of atoms set in keyword FIND. *Shake&wARP* by comparision : i) builds into much lower density (1.0s), ii) removes below 0.6s, iii) begins from six differently and optimally perturbed starting models, where 10% of atoms have been randomly removed, and iv) perturbs remaining coordinates by an average of 0.25Å r.m.s.d. (Figure 1). The process of starting with different models and building into relatively low solvent density, followed by weighted map averaging, can effectively be viewed as a real space solvent flattening procedure which significantly increases map contrast. Density features contained in each map amplify, and the noise density represented by variably placed atoms will be effectively averaged out. The power of map averaging for phase improvement has been well established in cross-crystal form and NCS averaging techniques (Kleywegt and Read, 1997), and is one of the major reasons for increased clarity and contrast in *S&W* maps compared to those reconstructed by other techniques.

Convergence and model perturbation in dummy atom refinement. As the only initial information available to *Shake&wARP* originates from the input model phases and the diffraction data, a balance must be maintained between data quality and staring phase quality, beyond which the *ARP/REFMAC* cycles will fail to converge or to provide phase improvement. An estimate for the minimum resolution (better than 2.4 Å) for the applicability of *ARP*, has been provided (Perrakis et al., 1997) and it was noted that the higher the resolution, the better the method will work (subject to other omni-valid criteria such as data quality and completeness). Need for convergence also affects the amount and mode of permissible coordinate perturbation. To investigate both the effect of various model perturbation methods as well as to provide an estimate for convergence of *S&W*, we calculated deviations between initial and perturbed models vs. the final structure, represented error as R-value and phase error in variation with resolution (figure 1). As expected, the higher the resolution of available data, the more robust the protocol becomes when starting from weak initial models, and given high resolution data, the capability of *S&W* to extend phases from marginal models having essential random phases at higher resolution is remarkable.

Structure factors and phase errors up to 1.2 Å for variously perturbed models used in the structure solution of a cytochrome c' dimer (RSCP, PDB entry 1GQA) from *Rhodobacter sphaeroides* (Ramirez et al., 2003) were calculated and results are explained in Figure 1. In summary, Gaussian (error function, graph B) perturbation alone tends to have little effect at low resolution (compared to deletion alone, graph A) while at the same time it introduces overly large phase errors at higher resolution. A simple random perturbation (range 0 – 0.5 Å), combined with 10% random atom deletion (D) yields a smoothly

increasing phase perturbation over the whole resolution range. A second set of graphs in Figure 1 shows R-value and phase error between (correctly placed) initial model 1CPQ and the final structure (1GQA). After the first round of rebuild, the phase error between the first rebuild and final structure is comparable to the phase error introduced by perturbation, indicating to what significant degree any starting model becomes perturbed. As would be expected in the case of bias removal, the correlation between the final model 1GQA and the *S&W* map from the first rebuild (0.89) is much better than the correlation between *S&W* map from first rebuild and first rebuild model itself (0.76).

For poor MR solutions (models with an initial correlation coefficient (CC) in the range 0.3-0.35) automated sequence correction and an initial step of *CNS* slow cool simulated annealing and torsion refinement (Adams et al., 1999) can be used to assure convergence. The weakest MR model we have been able to rebuild using *S&W* and manual rebuilding had a CC of 0.32 (Rv3465, 1.6 Å data, Figure 2). The absolute value of the CC, however, critically depends on the quality of the low resolution data (Dauter and Wilson, 2001), and does not necessarily provide a good predictor for the convergence of the *S&W* procedure against high resolution data. For the purpose of bias removal and/or structure validation, a single run of *S&W* is sufficient to provide reliable local analysis by real space correlation plots described in the corresponding section.

Implementation of the S&W bias removal service

Although the interactive interface of *CCP4* (Potterton et al., 2003) allows even novice users to navigate through the *CCP4* program suite with relative ease, the scripting of a complex, distributed routine like *Shake&wARP* would be a daunting task. We

therefore have implemented *S&W* as a easy to use web service incorporating several utility routines which clean up and convert the input coordinate files, standardize the PDB model file, and select proper parameters for the about 20 different *CCP4* programs called. We also incorporated run-time routines to produce GIF format plots of real space correlation and B-factor plots, B-factor histograms and, provided intensities were supplied, Wilson plots. The web routine provides the most common options, but does not allow for extensive experimentation with all parameters. The set of parameters we have selected as defaults are based on significant experience with the program on several platforms, and work for the vast majority of cases. Stability and load limitations of the web service impose certain limitations on parameter choice.

The web service (http://tuna.tamu.edu) was implemented on a Linux cluster (RedHat Linux 7.3) controlled by one four CPU main web and Condor job queuing server, and six dual CPU nodes execute the jobs distributed via Condor version 6.2.2. The web server scripts were written using Perl-Cgi and Perl v5.6.1 as the scripting language. When a job is submitted through the web server, a validation program (F90, described below) is executed, and a shell script splits the submission into 6 parallel sub-jobs in the Condor queue sending the jobs to free CPU nodes. The queue control then waits for all 6 jobs to complete and the main server continues post-processing and finalizing the output. The general program flow is depicted in Figure 3.

Input preparation. Only two data files and a few control parameters need to be provided to start *S&W*, the model in PDB format and reflection data in *Scalepack* (Otwinowski and Minor, 1997), *XPLOR/CNS* (Brünger et al., 1998) or ASCII text format. Other required input includes cell constants (if data are not in *Scalepack* format); number of molecules

per asymmetric unit and the number of residues pre asymmetric unit (used to determine the optimum number of atoms to be placed/removed in each *arp_waters* cycle, and for F_{ooo} estimates in truncate). Following control parameters are selectable: Optional molecular replacement and multi-segment rigid body refinement with *EPMR* (up to 3 molecules per a.u. are allowed); use of poly-alanine model which can be advantageous for sharpness of *EPMR* solutions (Kissinger et al., 2001); removal of water atoms (automatically if MR was selected); the standard option of executing bias removal and creating plots. Figure 4 shows the simple input panel of the server front page.

User data and control selections are initially checked for consistency at the client level using java scripts, and after the initial input validation, a preparation routine checks data for consistency, prepares and converts data files into *CCP4* format, and checks and standardizes the PDB file. A number of additional control parameters are derived from the user input by the setup routine and written to a project file. In stand-alone version this project file can be edited in order to perform special tasks with non-standard parameter combinations. The input validation program derives the following *CCP4* settings: resolution limits, FFT grid spacings according to resolution and space group, limits for *fft*, *arp_waters*, and *sfall*, and the number of atoms to remove/rebuild, according to asymmetric unit cell contents. In the web implementation, the number of *S&W* cycles is fixed at 30, although the slope of the R-value convergence is reported and allows automated termination.

A report of the check and the parameter settings is returned to the web client, and input can be corrected, or execution of the initial script is started. After further data preparation and standardizing of the files, optional MR is executed on the main server (20

cycles of *EPMR* if convergence is not reached, 12-4 Å data), followed by multi-segment rigid body refinement against data to 2.8Å. Subsequently, six scripts are generated and submitted to the Condor queuing system.

Output of Results. After initial validation, the user is prompted whether to continue executing the job. Once user confirms, an e-mail notification is sent, and it takes one to twenty hrs depending on the complexity of the problem and server workload for results to become available. In the meantime, the user is able to see (or download, if curious) the temporary files and logs while the job is progressing. If the job succeeds, an e-mail notice is sent and the following can be retrieved from the server via the results web page:
A *.phs file (h, k, l, F, FOM, PHWT) to create a bias minimized map with f*fom and phwt as Fourier coefficients (best viewed in *XtalView/Xfit* (McRee, 1999) by selecting f*fom as map type).
The re-placed input model, if MR was selected.
For each chain in the model, publication quality GIF images of real space correlation plots between the fit of the model to the electron density combined with per-residue B-factor plots as well as s cumulative B-factor histogram (examples given in the subsequent sections).
The data file converted into *.sca, *.fin, and mtz format (5% free flags set).
The phased data in MTZ format (HKL, FP, SIGFP, FOM, PHWT, FreeR_Flag). The free set is used internally only to estimate s_A in the *REFMAC* ML dummy atom refinement, and except in the case of a new MR solution, one should continue using the original free set.

The website introduction page includes further detailed description regarding usage, file formats, control parameters, job control, interpretation of results, and licensing (only a *CCP4* license is required to use the service).

Real space correlation plots

Global indicators of structural quality such as R-value and freeR (Brünger, 1992) convey very little about the actual correctness of the structure, and numerous examples exist of partially (or purposefully for demonstration) incorrectly traced structures with unsuspicious statistical descriptors (Dodson et al., 1996, Kleywegt and Jones, 1995). In the case of molecular replacement structures, even checks based on plausibility of the local geometry such as implemented in *WHAT-IF* or *PROCHECK* may not immediately trigger strong warning signs, in particular at low resolution and when refined with molecular dynamics protocols, where geometric restraints dominate the refinement (Dodson et al., 1996). In general, careful inspection of regions flagged in geometry checks, particularly Ramachandran plots (Sasisekharan, 1962, Ramachandran et al., 1963), nearly always reveals problems with a structure (Kleywegt and Brünger, 1996, Rupp and Segelke, 2001). However, the most comprehensive and fastest assessment of local quality – provided structure factor amplitudes are available - is the real space (RS) correlation coefficient (CC) between the calculated model map and the 'experimental' map calculated from observed intensities (Branden and Jones, 1990), particularly when the map contains a minimum of model bias. The RSCC has the benefit of being scale independent compared to real space R-values, and atoms placed correctly in weak density still correlate highly. Areas with low real space correlation coinciding with areas of high B-factors indicate that model tracing in these areas is in all likelihood genuinely

ambiguous due to lack of electron density. Deviations from the anti-correlation of B and RSCC nearly always indicate problem areas worth investigating (examples are provided in the next section). *SFCHECK* (Vaguine et al., 1999) and *mapoverlap* from the *CCP4* suite provide real space correlation analysis. From a survey of the literature, however, it appears that RSCC plots are not as frequently used as they probably should be. The electron density server (EDS) at the University of Uppsala (http://portray.bmc.uu.se/eds/) is a very useful web tool to locate potential problem areas in deposited structures. Such analytical web tools can be further enhanced to allow users to submit their coordinates and structure factor files. Application of map improvement and phase bias reduction routines such as the *Shake&wARP* service with return of corresponding RSCC plots and weighted Fourier map coefficients to the submitter for further refinement and rebuilding would probably promote the use of RSCC plots and contribute to increasing the quality of deposited structures.

Map improvement and bias reduction at work

The following section provides examples where *Shake&wARP* maps as produced by the web service. Examples include map improvement at various resolution, state of completeness, and reconstruction of absent parts or removal of questionable model parts or ligands. Even less spectacular improvements in map quality can make the difference between a clearly traceable map and a frustrating refinement stalled at high R values, in particular for less experienced model builders, who then are more likely to succeed and to avoid some of the mishaps we show in the examples. The clarity of the averaged maps

obtained from nearly finished models allows unambiguous identification of ligands and detailed fine-tuning of structural models.

Model correction and improvement

Sequence correction at 1.8Å in cytochrome c' from *Rhodobacter sphaeroides*. In the crystal structure solution of a cytochrome c' dimer (1GQA) from *Rhodobacter sphaeroides* (Ramirez et al., 2003), initial phases were obtained from a modest MR solution (CC=0.44 after rigid body refinement) using *EPMR* (Kissinger et al., 1999) with the coordinates of the *Rhodobacter capsulatus* cytochrome c' (1CPQ) as a search model (Tahirov et al., 1996). After the first round of sequence adjustment during model building into maps generated by *Shake&wARP*, a mismatch of the sequence became evident (Figure 5). Note that the most significant improvement occurs after averaging of the six *Shake&wARP* runs, attesting to the power of map averaging for density improvement (Kleywegt and Read, 1997). It must be also noted that improvements over the *REFMAC* ML coefficient map come at a substantial price in computational effort - *Shake&wARP* spawned a total of 150 runs of unrestrained *REFMAC* ML refinements.

The elusive N terminus of Calmodulin, 1.8 Å. In a near final model of Calmodulin, the N-terminal three residues could not be unambiguously built into *CNS* simulated annealing omit maps (provided by R. Skeene and B. Phipps, unpublished). When this model was subjected to a full bias removal run (automated MR using *EMPR* followed by *Shake&wARP*), the correct connecting electron density became clearly visible (Figure 6B), and the previously unmodelled, initial three residues could be unambiguously built backwards from the fourth residue (Figure 6D). While an experienced model builder

might recognized the missing residues, the incorrect connectivity apparent in the SA-omit map (Figure 6A) is likely to complicate model building in the N-terminal region.

Low resolution data

Apolipoprotein E4. The applicability of the *ARP* procedure, which in turn determines the resolution limit past which *Shake&wARP* can be used, has been discussed in detail (Lamzin and Wilson, 1993). Subject to effects of map noise, data completeness, and other effects that impair map quality, even a reasonable 2.5 -2.8 Å MR model, however, should allow the application of *Shake&wARP*. Apolipoprotein E4 (ApoE4) was solved, fully automated, from an ApoE3 search model 1BZ4 (Segelke et al., 2000) using *EPMR*, followed by rigid body refinement against the 2.5 Å data set, and *Shake&wARP*. Unambiguous visibility of the ApoE3/ApoE4 isoform difference (C112R) between ApoE3 (model) and ApoE4 (electron density) is evident even at a resolution approaching the limit of applicability of the underlying *ARP* program (Figure 7).

LysA from *M. tuberculosis*. LysA is an essential gene of *M.Tb*. involved in the last step of lysine biosynthesis through stereospecific decarboxylation of meso-diaminopimelic acid (Gokulan et al., 2003). Strong data to 2.8 Å were available, and the initial protein only structure model was submitted to the web service. The resulting averaged electron density (Figure 8A) clearly showed soaked PLP (vitamin B_6) covalently bound as the cofactor, and the product lysine (added in excess to crystallization cocktail). A further low-resolution example is provided below (BABIM complex).

(Un)ambiguous ligands

Uma. In a 2.2 Å structure of PcaA, an S-adenosyl-L-methionine dependent methyltransferase from *Mycobacterium tuberculosis* (Huang et al., 2002), we demonstrate the capability of *Shake&wARP* to recover ligands in complex structures. The presumed ligand S-adenosyl-L-homocysteine (SAH) was excluded from the model, and the remainder of the model submitted to the TB Bias Removal Server. The *Shake&wARP* map in Figure 8B clearly recovers the SAH ligand.

Clostridium Botulinum Serotype B Neurotoxin light chain protease - BABIM complex. The BABIM complex of *Clostridium botulinum* neurotoxin serotype B light chain (BotLCB) protease (Hanson et al., 2000) was submitted to the web service. The data reportedly extend to only 2.7 Å, but 2.5Å data were deposited in the PDB (entry 1FQH), and these were used without any sigma cutoffs for *Shake&wARP*. As an additional control for recovery of electron density, a residue close to the BABIM inhibitor (E170, B=38 $Å^2$) and the catalytic Zn atom were also removed. The *Shake&wARP* map in Figure 9 clearly recovers the omitted residue (perhaps not quite unambiguously built), as well as the Zn atom, and the oxygen atom of the catalytic water. However, despite 0.5 s map contouring, there is no indication of the BABIM ligand. Given its reported excessive average B-factors of 130$Å^2$, the inhibitor BABIM, exhibiting scarcely few contacts to the protease, unfavorable geometry, and little if any electron density, is not likely to be present in any substantial amount this structure. Based on these findings, a correction has been published (Hanson et al., 2002).

Clostridium Botulinum Serotype B Neurotoxin light chain protease - Synaptobrevin-II complex. A dramatic example of where the use of a real space

correlation plot would have provided early warning signs of an incorrect model is the complex of BotLCB with synaptobrevin (1F83). The plot created by the web service (Figure 10) reveals extremely poor real space correlation and excessive B-factors for the ligand. Severe problems with the ligand refinement, including absence of the ligand, must be expected. It is worthwhile mentioning that deposition of structure factors for both BotLCB complexes indicates that an honest mistake was made. Suppression of structure factors, when obvious warning signs are present, may shed serious doubt on the validity of a structure, as recently discussed (Kleywegt and Jones, 2002).

Buffers make excellent ligands. A taste receptor binds a number of molecules, some of which taste up to 20,000 times as sweet as sugar (K.Gokulan, unpublished). A refined and completed model of the structure with omitted ligand was submitted to *Shake&wARP*, and the resulting map shed serious suspicion about the presence of the ligand. While the *CNS* SA-omit ML map may have sufficed to convince the proficient crystallographer that the ligand modeling was dubious (Figure 11A), the enhanced clarity of the *Shake&wARP* map proves it beyond doubt (Figure 11B) - thus overcoming even wishful mental bias. Based on electron density the ligand was identified as a sulfonate buffer. Subsequent consulting of the crystallization protocol confirmed the presence of zwitterionic TES buffer (2-[2-Hydroxy-1,1-bis(hydroxymethyl)ethyl-amino]ethanesulfonic acid), and the corresponding structure has been modeled in the density.

Conclusions

Consistent use of map validation tools, including real space correlation plots, can prevent the great majority of bias-caused errors commonly found in crystallographic models. Although these validation methods exist, and some have been introduced more than a decade ago, they are not as widely used as one would expect. We hope that public availability of our web service would make it convenient to use structure factor based validation techniques and thus contribute to increased quality of protein structures. The concerning trend for structure factors to be absent when global quality indicators are poor has been pointed out recently (Kleywegt and Jones, 2002), and we hope also that deposition of structure factors and their use for structure validation become prevailing practice, as has been the case in small molecule crystallography for many years. (At the time of this writing, less than 50% of deposited coordinates in the Protein Data Bank are accompanied with corresponding structure factor entries).

Acknowledgements

We thank the laboratory members of Jim Sacchettini and Joel Sussmann for kindly supplying many coordinate and data sets for blind test cases and for evaluating the *Shake&wARP* results. BR wishes to thank LLNL and James Sacchettini for supporting his sabbatical leave at Texas A&M University. Lawrence Livermore National Laboratory is operated by the University of California under contract W-7405-ENG-48 for the U.S. Department of Energy. This work was sponsored by NIH-NIGMS Grant No P50 GM62410 (TB Structural Genomics Consortium) and the Robert Welch Foundation at Texas A&M University.

References

Adams, P. D., Panu, N. S., Read, R. J. and Brünger, A. T. (1999) Extending the limits of molecular replacement through combined simulated annealing and maximum-likelihood refinement. *Acta Crystallogr,* **D55,** 181-190.

Bath, T. N. (1988) Calculation of an OMIT map. *J Appl Cryst,* **21,** 279-281.

Bath, T. N. and Cohen, G. H. (1994) OMITMAP: An eletron density map suitable for the examination of errors in a macromolecular model. *J Appl Cryst,* **17,** 244-248.

Blundell, T. L., Jhoti, H. and Abell, C. (2001) High-Throughput Crystallography for Lead Discovery in Drug Design. *Nature Reviews Drug Discovery*, **1,** 45-54.

Blundell, T. L. and Johnson, L. N. (1976) *Protein Crystallography,* Academic Press, London, UK.

Branden, C. I. and Jones, T. A. (1990) Between objectivity and subjectivity. *Nature,* **343,** 687-689.

Brünger, A. T. (1992) Free R value: A novel statistical quantity for assessing the accuracy of crystal structures. *Nature,* **355,** 472–475.

Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J. S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. and Warren, G. L. (1998) Crystallography & NMR System: A New Software Suite for Macromolecular Structure Determination. *Acta Crystallogr,* **D54,** 905-921.

CCP4 (1994) The CCP4 Suite: Programs for Protein Crystallography. *Acta Crystallogr,* **D50,** 760-763.

Dauter, Z. and Wilson, K. S. (2001) Principles of monochromatic data collection. *International Tables For Crystallography,* **F,** 177-195.

Dodson, E. J., Kleywegt, G. J. and Wilson, K. S. (1996) Report of a workshop on the use of statistical validators in protein x-ray crystallography. *Acta Crystallogr,* **D52,** 228-234.

Dong, L.-M., Wilson, C., Wardell, M. R., Simmons, T., Mahley, R. W., Weisgraber, K. H. and Agard, D. A. (1994) Human apolipoprotein E. *Role of arginine 61 in mediating the lipoprotein preferences of the E3 and E4 isoforms. J Biol Chem,* **269,** 22358–22365.

Gokulan, K., Rupp, B., Pavelka, M. S., Jr., Jacobs, W. R., Jr. and Sacchettini, J. C. (2003) Crystal structure of Mycobacterium tuberculosis diaminopimelate decarboxylase, an essential enzyme in bacterial lysine biosynthesis. *J Biol Chem,* **278,** 18588-18596.

Hanson, M. A., Oost, T. K., Rich, D. H., Stevens, R. C. and Sukonpan, C. (2002) Structural basis for BABIM inhibition of botulinum neurotoxin type B protease (Correction). *J Am Chem Soc,* **124,** 10248-10248.

Hanson, M. A., Oost, T. K., Sukonpan, C., Rich, D. H. and Stevens, R. C. (2000) Structural basis for BABIM inhibition of botulinum neurotoxin type B protease. *J Am Chem Soc,* **122,** 11268-11269.

Hanson, M. A. and Stevens, R. C. (2000) Cocrystal structure of synatptobrevin-II bound to botulinum neurotoxin type B at 2.0Å resolution. *Nature Struct Biol,* **7,** 687-692.

Hendrickson, W. A. and Ogata, C. M. (1997) Phase determination from multiwavelength anomalous diffraction measurements. *Meth Enzymol,* **276,** 494–516.

Hodel, A., Kim, S.-H. and Brünger, A. T. (1992) Model Bias in Macromolecular Structures. *Acta Crystallogr*, **48**.

Huang, C. C., Smith, C. V., Glickman, M. S., Jacobs, W. R., Jr. and Sacchettini, J. C. (2002) Crystal structures of mycolic acid cyclopropane synthases from Mycobacterium tuberculosis. *J Biol Chem,* **277,** 11559-69.

Islam, S. A., Carvin, D., Sternberg, M. J. E. and Blundell, T. L. (1998) HAD, a Data Bank of Heavy-Atom Binding Sites in Protein Crystals: a Resource for Use in Multiple Isomorphous Replacement and Anomalous Scattering. *Acta Crystallogr*, **D54,** 1199-1206.

Kissinger, C. R., Gehlhaar, D. K., Smith, B. A. and Bouzida, D. (2001) Molecular replacement by evolutionary search. *Acta Crystallogr*, **D57,** 1474-1479.

Kissinger, C. R., Gelhaar, D. K. and Fogel, D. B. (1999) Rapid automated molecular replacement by evolutionary search. *Acta Crystallogr*, **D55,** 484-491.

Kleywegt, G. J. and Brünger, A. T. (1996) Cross-validation in crystallography: practice and applications. *Structure,* **4,** 897-904.

Kleywegt, G. J. and Jones, T. A. (1995) Where freedom is given, liberties are taken. *Structure,* **3,** 535-540.

Kleywegt, G. J. and Jones, T. A. (1997) Model Building and Refinement Practice. *Meth Enzymol*, **277,** 208-230.

Kleywegt, G. J. and Jones, T. A. (2002) Homo crystallographicus - quo vadis? *Structure*, **10,** 465-472.

Kleywegt, G. J. and Read, R. J. (1997) Not your average density. *Structure,* **5,** 1557-1569.

Lamzin, V. S. and Wilson, K. S. (1993) Automated refinement of protein models. *Acta Crystallogr,* **D53,** 448-455.

Laskowski, R. A., MacArthur, M. W., Moss, D. S. and Thornton, J. M. (1993) PROCHECK: a program to check the stereochemical quality of protein structures. *J Appl Cryst,* **26,** 283-291.

Matthews, B. W. (2001) Heavy atom location and phase determination with single wavelength diffraction data. *International Tables for Crystallography,* **F,** 293-298.

McRee, D. E. (1999) XtalView/Xfit - a versatile program for manipulating atomic coordinates and electron density. *J Struct Biol,* **125,** 156-165.

Merritt, E. A. and Bacon, D. J. (1997) Raster3D: Photorealistic molecular graphics. *Meth Enzymol,* **277,** 505–524.

Murshudov, G. N., Vagin, A. A. and Dodson, E. D. (1997) Refinement of Macromolecular Structures by the Maximum-Likelihood Method. *Acta Crystallogr,* **D53,** 240-255.

Norvell, J. C. and Zapp-Machalek, A. (2000) Structural genomics programs at the US National Institute of General Medical Sciences. *Nature Struct Biol Suppl,* **7,** 931.

Otwinowski, Z. and Minor, W. (1997) Processing of X-ray Diffraction Data Collected in Oszillation Mode. *Meth Enzymol,* **267,** 307-326.

Perrakis, A., Sixma, T. K., Wilson, K. S. and Lamzin, V. S. (1997) wARP: Improvement and Extension of Crystallographic Phases by Weighted Averaging of Multiple-Refined Dummy Atomic Models. *Acta Crystallogr,* **D53,** 448-455.

Potterton, E., Briggs, P. J., Turkenberg, M. and Dodson, E. D. (2003) A graphical user interface to the CCP4 program suite. *Acta Crystallogr,* **D59,** 1131-1137.

Ramachandran, G. N., Ramakrishnan, C. and Sasisekharan, V. (1963) Stereochemistry of polypeptide chain configurations. *J Mol Biol,* **7,** 95–99.

Ramirez, L., Axelrod, H., Herron, S., Rupp, B., Allen, J. and Kantardjieff, K. A. (2003) High resolution crystal structure of ferricytochrome c' from Rhodobacter sphaeroides. *J Chem Cryst,* **33,** 413-424.

Read, R. J. (1986) Improved Fourier coefficients for maps using phases from partial structures with errors. *Acta Crystallogr,* **A42,** 140-149.

Read, R. J. (1990) Structure-factor probabilities for related structures. *Acta Crystallogr,* **A46,** 900-912.

Read, R. J. (1997) Model phases: probabilities and bias. *Meth Enzymol,* **278,** 110-128.

Rupp, B. (2003) High throughput crystallography at an affordable cost: The TB Structural Genomics Consortium crystallization facility. *Acc Chem Res,* **36,** 173-181.

Rupp, B. and Segelke, B. W. (2001) Questions about the structure of the botulinum neurotoxin B light chain in complex with a target peptide. *Nature Struct Biol,* **8,** 643-664.

Sasisekharan, V. (1962) *Stereochemical criteria for polypeptide and protein structures,* Wiley and Sons, Madras, India.

Segelke, B. W., Forstner, M., Knapp, M., Trakhanov, S. D., Parkin, S., Newhouse, Y. M., Bellamy, H. D., Weisgraber, K. H. and Rupp, B. (2000) Conformational flexibility in the Apolipoprotein E amino-terminal domain structure determined

from three new crystal forms : Implication for lipid binding. *Prot Science,* **9,** 886-897.

Tahirov, T. H., Misaki, S., Meyer, T. E., Cusanovich, M. A., Higuchi, Y. and Yasuoka, N. (1996) High-resolution crystal structures of two polymorphs of cytochrome c' from the purple phototrophic bacterium Rhodobacter capsulatus. *J Mol Biol,* **259,** 467-479.

Terwilliger, T. C. (1999) Reciprocal space solvent flattening. *Acta Crystallogr,* **D55,** 1863-71.

Terwilliger, T. C. (2000) Maximum likelihood density modification. *Acta Crystallogr,* **D56,** 965-972.

Vaguine, A. A., Richelle, J. and Wodak, S. J. (1999) SFCHECK: a unified set of procedures for evaluating the quality of macromolecular structure-factor data and their agreement with the atomic model. *Acta Crystallogr,* **D55,** 191-20.

Vriend, G. (1990) WHAT IF: A molecular modeling and drug design program. *J Mol Graphics,* **8,** 52–56.

Zeng, Z.-H., Castaño, A. R., Segelke, B. W., Stura, E. A., Peterson, P. A. and Wilson, I. A. (1997) Crystal structure of mouse CD1: An MHC-like fold with a large hydrophobic binding groove. *Science,* **277,** 339-345.

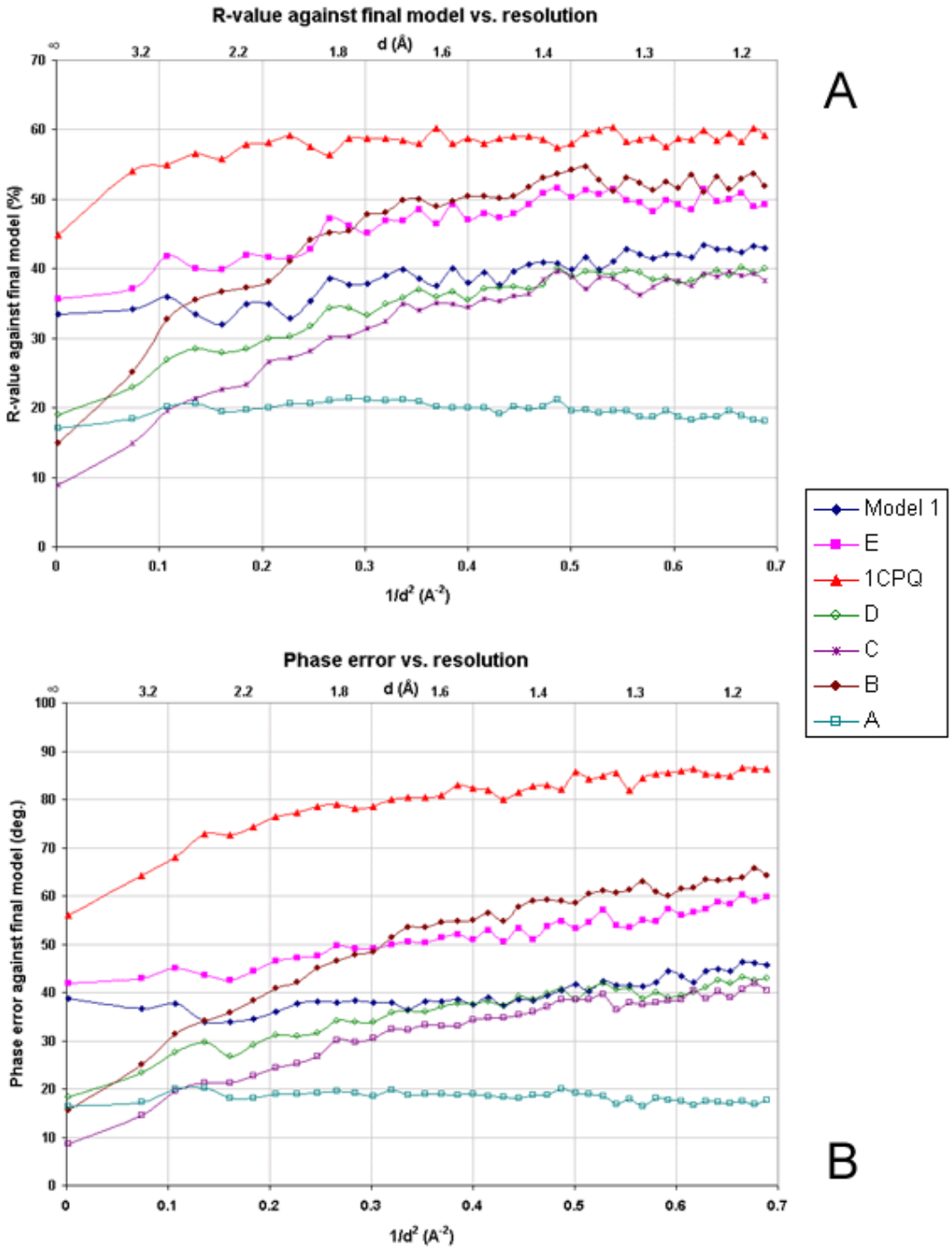

Figure 1. R-value (top panel) or phase error (bottom panel) between final CCP structure model (1GQA) and varying models used in *S&W*. Graph legends are as follows. **1CPQ**: the re-placed initial MR model; **Model 1**: sequence corrected and CNS

simulated annealing torsion angle refinement of 1CPQ model; **A**: 10% atoms of final model deleted at random; **B**: Error function perturbation of final model with coordinate deviation s of 0.25 Å ; **C**: Linear random perturbation of final between 0 and 0.5 Å (mean = 0.25Å); **D**: 10% random atom deletion and linear random perturbation between 0 and 0.5 Å; **E**: Model 1, perturbed as in **D**. While error function perturbation mode **B** alone yields very high perturbation at higher resolution and has little effect at low resolution, combination **D** appears to be an optimal compromise and yields smoothly increasing model perturbation with increasing resolution. Phase error in bottom panel given as figure of merit, f.o.m. = cos(?f). While at 12-3 Å 1CPQ is accurate enough as a model to yield a weak but clear MR solution, phases for the 1CPQ model are practically random beyond 3-2.5 Å, despite only 1.44 Å Ca r.m.s.d. for alignment of 1CPQ with the final model. Note that good experimental MAD/SAD phases have f.o.m.s approaching 0.8 - 0.9 even at high resolution (better than 2.0 Å), which emphasizes how weak and biased MR phases are in comparison.

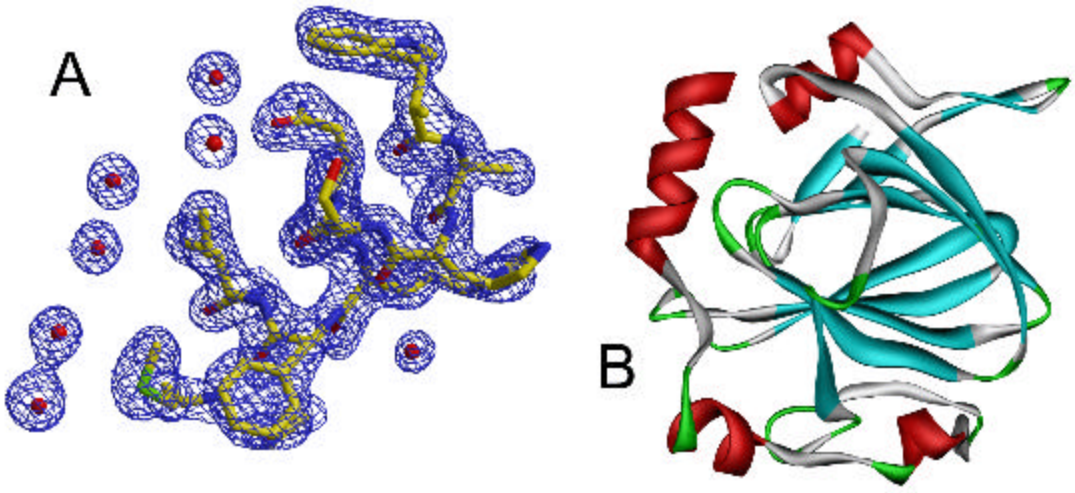

Figure 2. *M. tuberculosis* Rv3465 protein structure. Rv3465, a dTDP-4-dehydrorhamnose 3,5-epimerase, was the first structure entirely processed by facilities of the TB Structural Genomics Consortium using the *Shake&wARP* protocol from a poor staring model with an *EPMR* solution correlation coefficient of 0.32. Left panel: bias reduced *Shake&wARP* map contoured at one s electron density level showing the clarity of map and solvent definition. Final model of missing regions (not used in map calculations) is superimposed on map. Right side: ribbon diagram of final molecular structure. Intermediate steps of map improvement and comparison with *REFMAC* $2mF_o$-DF_c maximum likelihood maps are shown in (Rupp, 2003). All figures showing electron density have been created with *XtalView/Xfit* (McRee, 1999) and *Raster3D* (Merritt and Bacon, 1997).

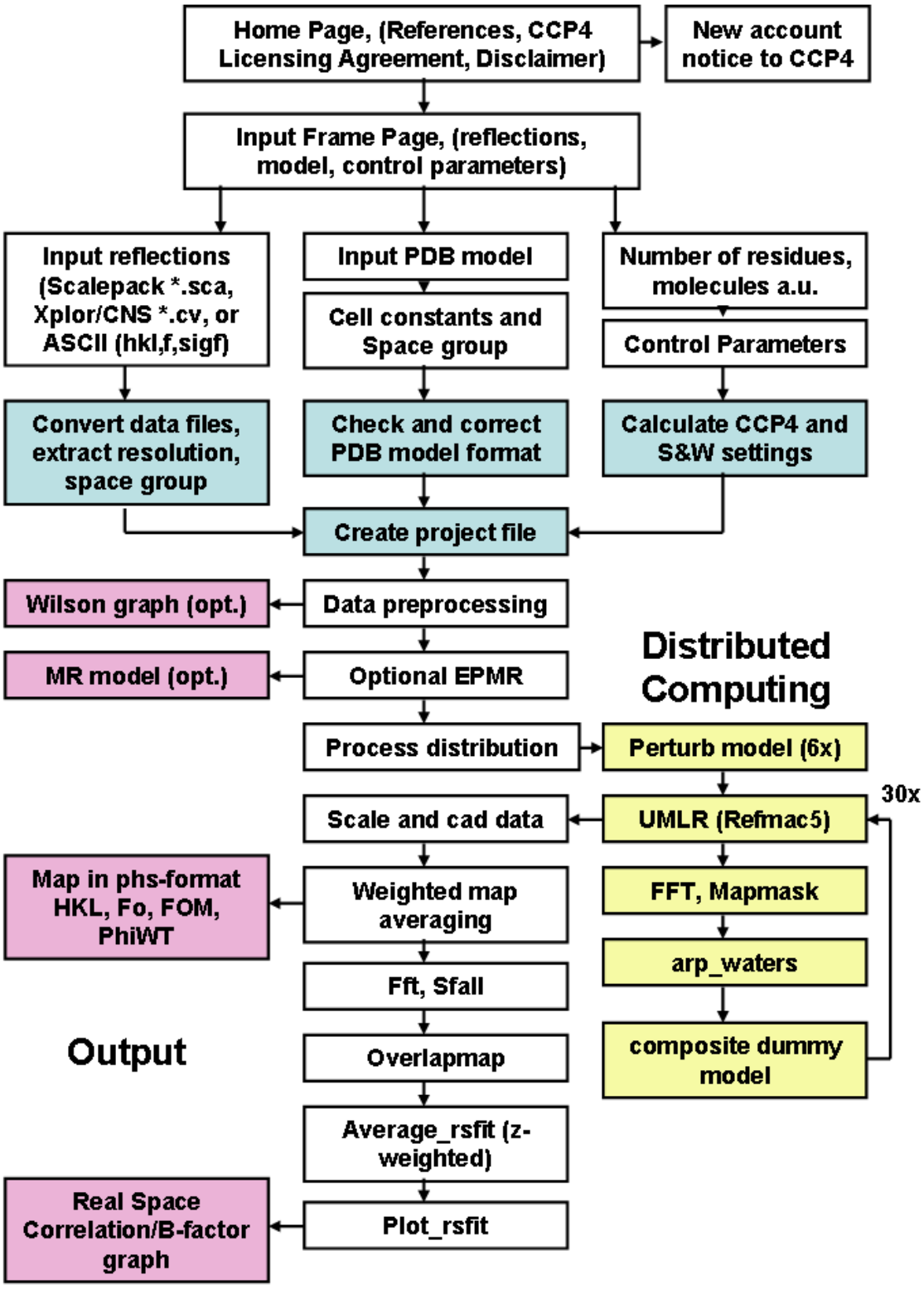
Home Page, (References, CCP4 Licensing Agreement, Disclaimer)
New account notice to CCP4
Input Frame Page, (reflections, model, control parameters)
Input reflections (Scalepack *.sca, Xplor/CNS *.cv, or ASCII (hkl,f,sigf)
Input PDB model
Number of residues, molecules a.u.
Cell constants and Space group
Control Parameters
Convert data files, extract resolution, space group
Check and correct PDB model format
Calculate CCP4 and S&W settings
Create project file
Wilson graph (opt.)
Data preprocessing
MR model (opt.)
Optional EPMR
Distributed Computing
Process distribution
Perturb model (6x)
30x
Scale and cad data
UMLR (Refmac5)
Map in phs-format HKL, Fo, FOM, PhiWT
Weighted map averaging
FFT, Mapmask
arp_waters
Fft, Sfall
Output
Overlapmap
composite dummy model
Average_rsfit (z-weighted)
Real Space Correlation/B-factor graph
Plot_rsfit

Figure 3: Schematic program flow of a *S&W* web submission. Blue, initial input preparation and validation; yellow, iterative steps conducted on cluster server members; magenta, output files

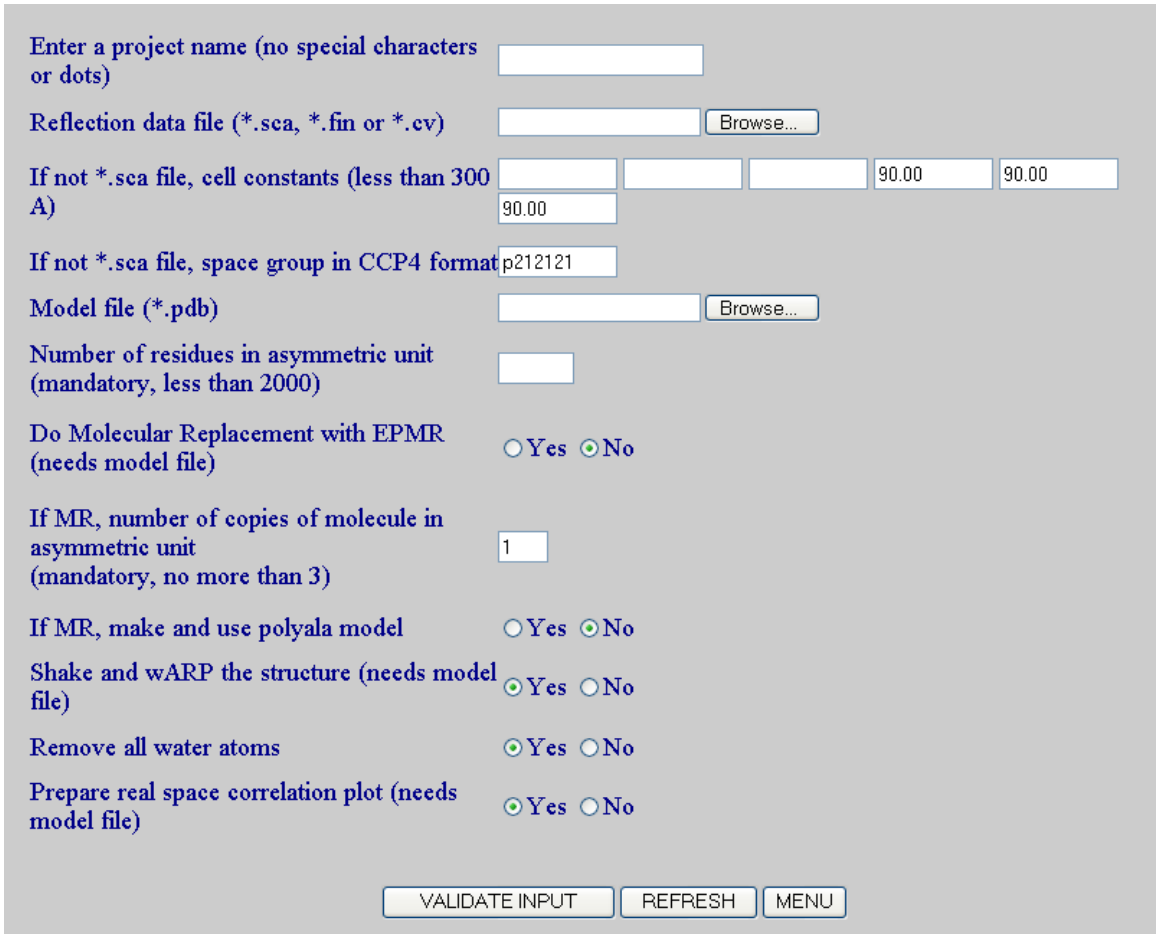

Figure 4: The simple user interface of the TB consortium bias removal service. All other program parameters are calculated from the input data, and if the validation results are consistent, the user can submit the job.

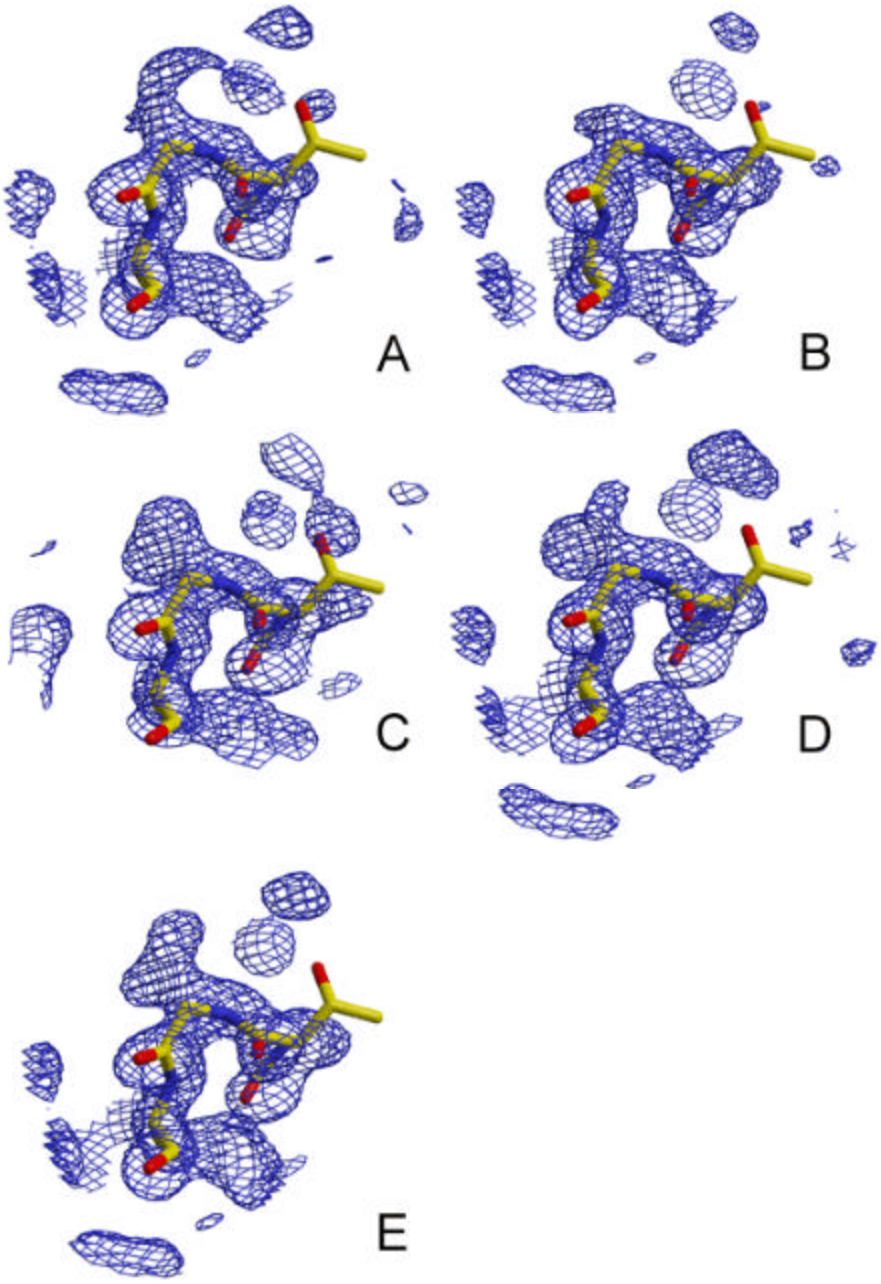

Figure 5. Sequence error in RSCP identified by electron density. Amino acid stretch 116GTGC119 should be GGTC. 1.8Å data, varying map types, 1 s electron density level

in blue. (A), plain 2Fo-Fc map; (B) *REFMAC* ML map (2mFo-2DFc); (C) single *ARP*-type map run 1; (D) single *ARP*-type map run 2, (E) combined *Shake&ARP* map. (E) shows the correct electron density most clearly. The same BLOB settings in *Xfit* have been used to display the maps in vicinity of the residues, no other editing (or 'density modification') tool has been used.

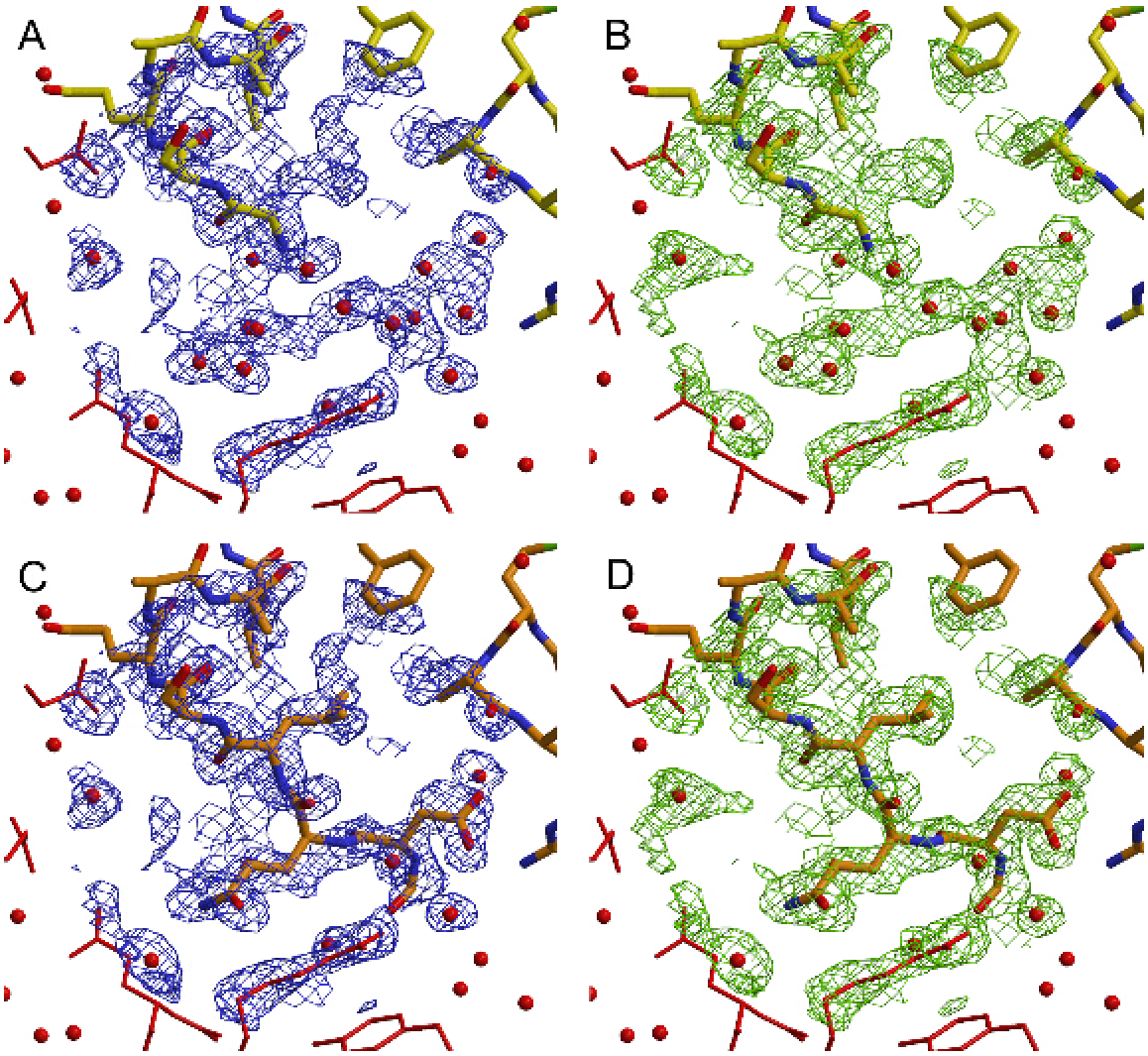

Figure 6. **Corrected connectivity for three N-terminal residues in Calmodulin**. (A) Incomplete model placed in *CNS* 2Fo-Fc map (blue contours, 1s). (B) Incomplete model placed in *Shake&wARP* map (green contours, 1s). (C) Complete model fit to *CNS* 2Fo-Fc electron density. (D) Completed model fit into *Shake&wARP* electron density. Although the N-terminal three residues could likely have been placed into the *CNS* 2Fo-Fc electron density by an experienced crystallographer, the incorrect map connectivity before residue Leu 4 and would likely stall automated chain tracing and model building

(or a less experienced model builder). The correct connectivity of electron density is quite clear in the *Shake&wARP* map.

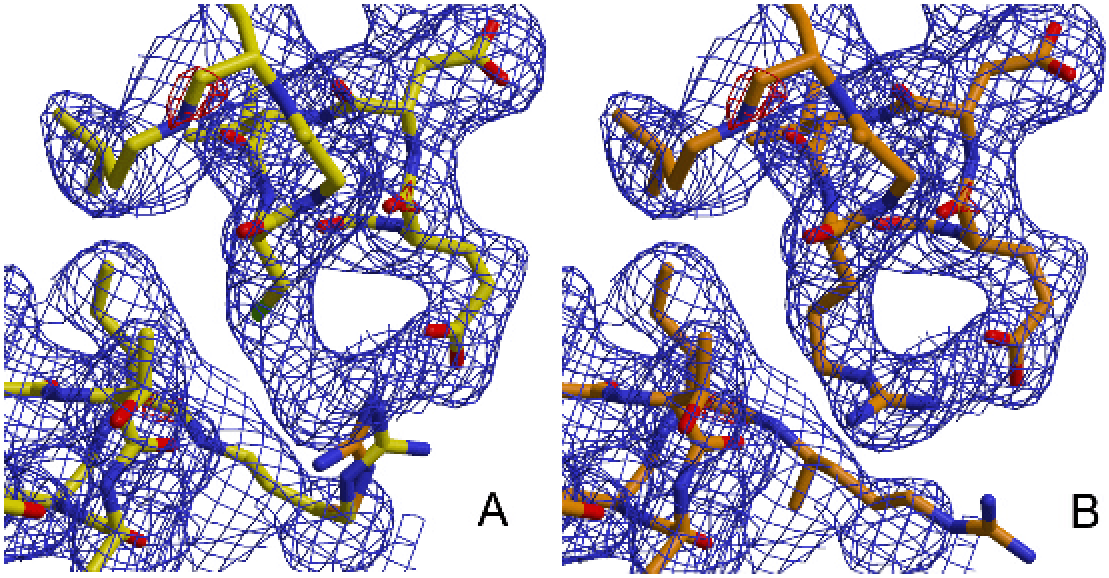

Figure 7. ApoE3/4 isoform difference C112R. Density shown is that for the ApoE4 isoform, solved by MR in a new crystal form (B. Rupp, unpublished data). The ApoE3 model (112C) is shown in panel (A). In the E3 isoform, 109E is hydrogen bonded to 61R. ApoE4 model is shown in panel (B). 112R clearly fits into electron density, making a new hydrogen bond to 109E, and disrupting the hydrogen bond to 61R. 61R adopts an extended conformation, changing the charge disposition of the helix 2- helix 3 surface and affecting VLDL binding (Dong et al., 1994).

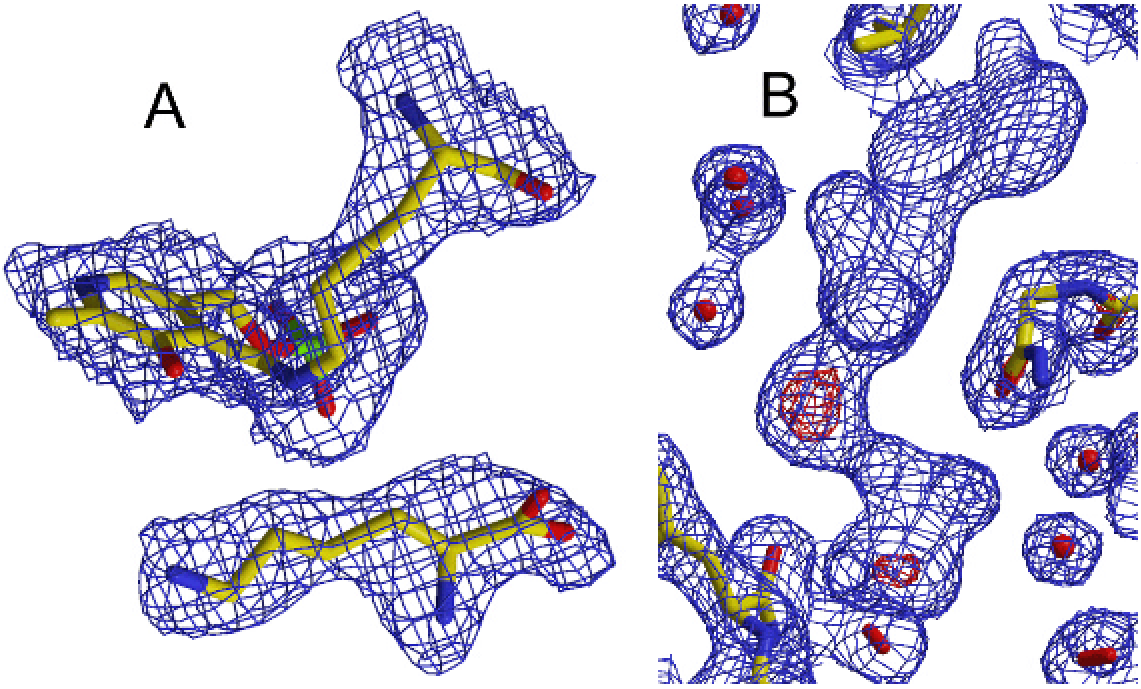

Figure 8. *Shake&wARP* map of *M. tuberculosis* drug target structures. Left panel: covalently bound cofactor PLP and product lysine clearly recovered in the active site of lysA, a diaminopimelate decarboxylase, at 2.8Å (Gokulan et al., 2003). Right panel: Right panel: PcaA, an S-adenosyl-L-methionine dependent methyltransferase from *Mycobacterium tuberculosis.* The standard *Shake&wARP* protocol at 2.2 Å clearly recovers the electron density of the S-adenosyl-L-homocysteine (Huang et al., 2002).

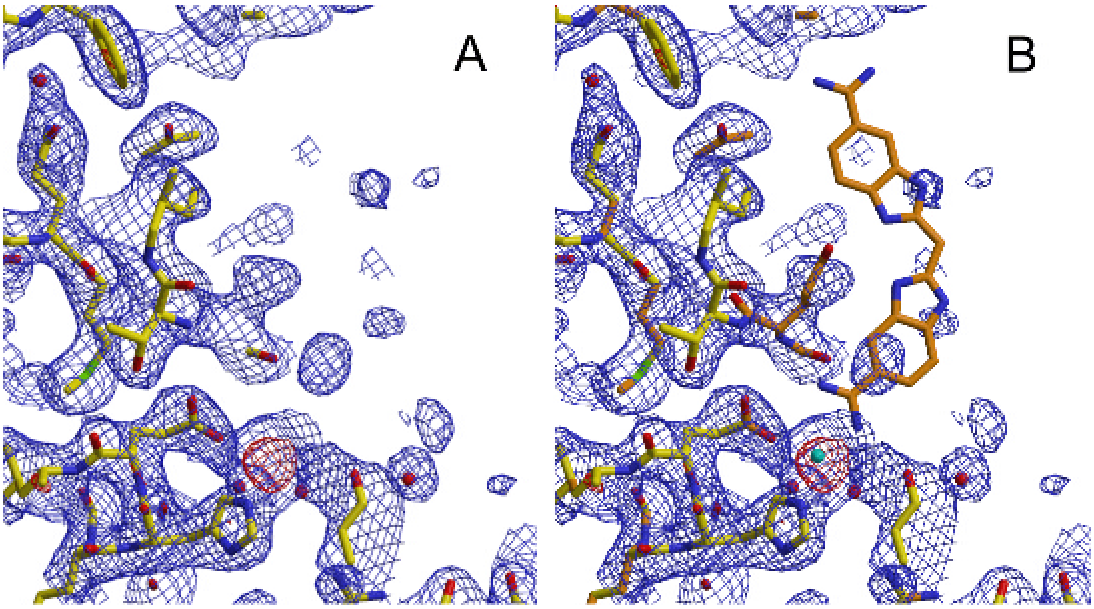

Figure 9. Clostridium Botulinum Serotype B Neurotoxin light chain protease - BABIM complex. This structure 1FQH (Hanson et al., 2000) was subjected to the standard *Shake&wARP* procedure. Data were used as deposited (2.5 Å), without any sigma cutoffs. BABIM, E170 and catalytic Zn ion were omitted. E170 and the Zn ion are clearly recovered by the *Shake&warp* procedure. However, little if any electron density is evident for the planar BABIM ligand, despite 0.5 s map contouring.

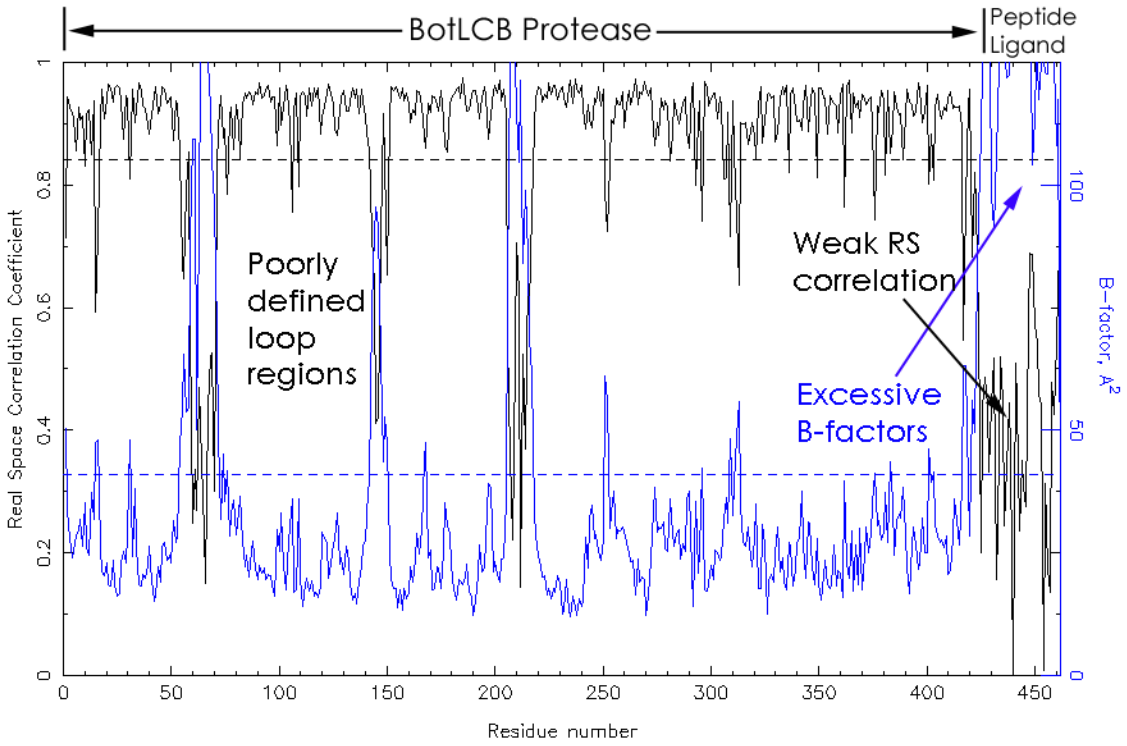

Figure 10: Real space correlation coefficient and B-factors plot. PDB entry 1F83 contains the model coordinates for the BotLCB protease - synaptobrevin-II complex (Hanson and Stevens, 2000). Shown in black (upper curve) is the residue-by-residue real space correlation coefficient, in blue the B-factors are plotted for each residue. The left part of the figure corresponds to the protease, which, with exception of a three loop regions shows normal behavior. The synaptobrevin-II ligand peptide at the right figure edge, however, shows a very worrisome cross-over between abysmal real space correlation and excessive B-factors. A simple plot of this nature, inspected beforehand or submitted with the manuscript, would have raised sufficient flags to prevent the public

discourse regarding the validity of the results (Rupp and Segelke, 2001). The plot (less descriptive labeling) was created by the *S&W* service.

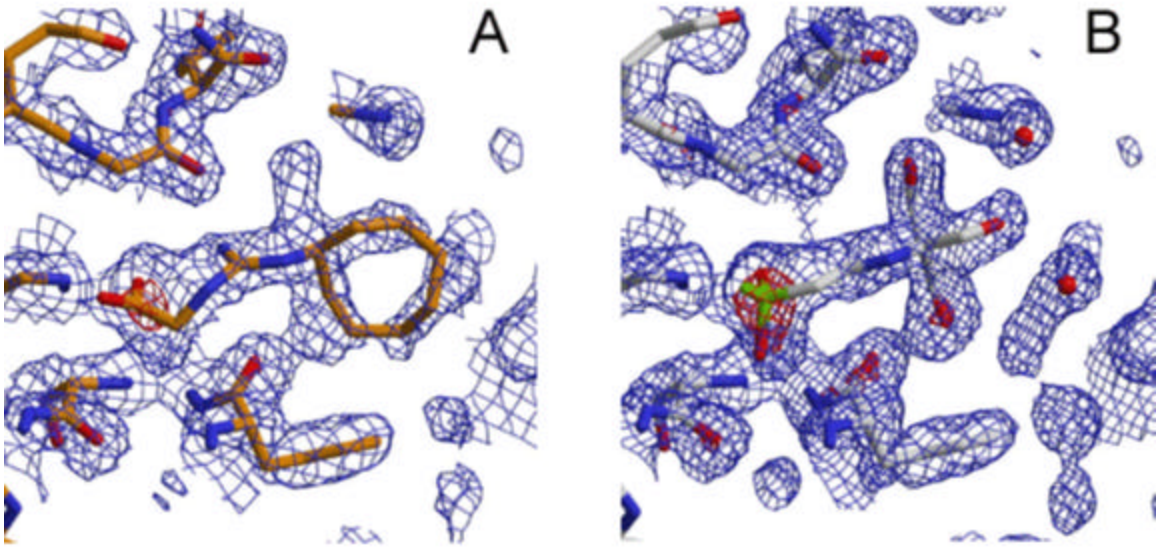

Figure 11. **TES buffer in ligand binding site**. Maps contoured at 1s (blue) and 5s (red). (A) presumed ligand built into *CNS* ML 2Fo-Fc map; (B) *Shake&wARP* map, with TES buffer built into density. Map has less noise and cleaner connectivity and reveals the true nature of the ligand. A questionable VdW contact is also obvious between 'ligand' and protein in the left panel (A).